Journal of the Korean Statistical Society 45 (2016) 33–50

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Non-parametric smoothed estimation of multivariate cumulative distribution and survival functions, and receiver operating characteristic curves

Tarn Duong *

Sorbonne Universities, University Pierre and Marie Curie—Paris 6, Theoretical and Applied Statistics Laboratory (LSTA), UR 1, F-75005, Paris, France

ARTICLE INFO

Article history:
Received 14 November 2014
Accepted 29 June 2015
Available online 19 August 2015

AMS 2000 subject classifications:
primary 62G05
secondary 62H10

Keywords:
Asymptotic mean integrated squared error
Diagnostic test
Optimal bandwidth matrix selection
Quantile
ROC curve

ABSTRACT

A unified framework to analyse multivariate kernel estimators of distribution and survival functions is introduced, before turning our attention to receiver operating characteristic (ROC) curves. These are well-established visual analytic tools for univariate data samples, though their generalisation to multivariate data has been limited. Since non-parametric multivariate kernel smoothing methods possess excellent visualisation properties, they serve as a solid basis for their estimation. With optimal data-based bandwidth matrix selectors, we demonstrate that they possess suitable properties for exploratory data analysis of simulated and experimental data.

1. Introduction

A basic problem in multivariate data analysis is estimating cumulative distribution functions, though there has been a relative paucity of their analysis as compared to density functions. The former are however important in a wide range of data analytic situations. We set up a unified framework to treat kernel estimators of the multivariate distribution functions and the closely related survival functions, since kernel estimators are widely used in non-parametric data smoothing, see Simonoff (1996) and Wand and Jones (1995) for an overview. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ be a d-variate random variable with distribution F and density f. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, then we define the cumulative distribution of $\boldsymbol{X}$ to be

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{w})\, d\boldsymbol{w}$$

where $\int_{-\infty}^{\boldsymbol{x}} d\boldsymbol{w}$ is an abbreviation of $\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} dw_1 \cdots dw_d$. The survival function is defined as $\bar{F}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} > \boldsymbol{x})$, complementary to the cumulative distribution function $F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x})$. The usual relation $\bar{F}(x) = 1 - F(x)$ for univariate data does not hold for multivariate data since the hyper-rectangles $\{\boldsymbol{w} \in \mathbb{R}^d : \boldsymbol{w} \leq \boldsymbol{x}\} \cup \{\boldsymbol{w} \in \mathbb{R}^d : \boldsymbol{w} > \boldsymbol{x}\} \neq \mathbb{R}^d$ in general.

These functions are brought together in the analysis of receiver operating characteristic (ROC) curves. ROC curves were introduced in the context of signal detection, e.g. Peterson, Birdsall, and Fox (1954), though they have been subsequently

* Correspondence to: Sorbonne Paris City, University Paris-North—Paris 13, Computer Science Laboratory (LIPN), CNRS UMR 7030, F-93430, Villetaneuse, France.
E-mail address: duong@lipn.univ-paris13.fr.

http://dx.doi.org/10.1016/j.jkss.2015.06.002
1226-3192/

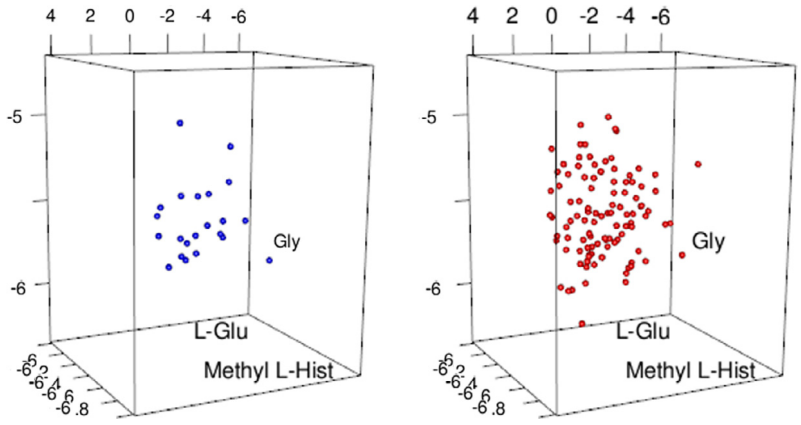

Fig. 1. Scatterplots for Spinal Muscular Atrophy (SMA) data set. The variables are the negative log concentrations (μM) of Glycine (Gly), L-glutamic (L-Glu) and 1-methyl-L-histidine (methyl L-Hist). (Left) Scatterplot for 22 age-matched healthy control children. (Right) Scatterplot for 108 SMA afflicted children.

widely adopted in many different contexts, whenever the values of a second population tend to be greater than those from the first one. The standard definition of ROC curves allows only for scalar valued diagnostic variables whereas the ability to handle multivariate data would be beneficial in many circumstances. Following Handcock and Morris (1998), Hsieh and Turnbull (1996) and Lloyd (1998), instead of comparing the vector of diagnostic variables to a threshold component-wise directly, we apply the survival function $\bar{F}_{\boldsymbol{X}_1}$ from the first population as a pre-transformation. This leads to us to define a multivariate ROC curve as the graph

$$\{(\mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_1) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x})), \mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_2) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}))) : \boldsymbol{x} \in \mathbb{R}^d\}.$$

We use this definition rather than the seemingly more straightforward generalisation from the univariate case $\{(\mathbb{P}(\boldsymbol{X}_1 > \boldsymbol{x}), \mathbb{P}(\boldsymbol{X}_2 > \boldsymbol{x})) : \boldsymbol{x} \in \mathbb{R}^d\}$ which is not a well-defined multivariate function, whereas the above ROC curve is monotonic by construction since it is a quantile–quantile plot. This approach is an alternative to current methods such as the dimension reduction via a weighted vector norm of Pepe and Thompson (2000) and Su and Liu (1993) or the singular value decomposition combined with likelihood ratios of Pfeiffer and Bura (2008); or the logistic regression modelling of Pepe (1998). The reader interested in a more comprehensive review is invited to consult Shapiro (1999) and the references contained therein. One of the main advantages of our proposed approach is that it does not require parametric assumptions on the underlying random variables or on the transformation, and so is an ideal candidate within a non-parametric smoothing framework.

A data set for which a multivariate ROC curve analysis would be beneficial is the Pilot Study of Biomarkers for Spinal Muscular Atrophy (BforSMA), available from http://neuinfo.org/smabiomarkers. The full database contains a large variety of measurements taken from a cohort of 130 children aged between 2 and 12 years, with 108 children with genetically confirmed Spinal Muscular Atrophy (SMA) and 22 aged-matched healthy controls. We take a subset of these data identified in Finkel et al. (2012), as potential biomarkers for SMA, namely the (negative log) concentrations of the amino acids Glycine (Gly), L-glutamic (L-Glu) and 1-methyl-L-histidine (methyl L-Hist). The 3-dimensional scatterplots in Fig. 1 give a visual impression that the point cloud of SMA patients have generally higher biomarker values than the control patients, and a ROC curve analysis would visualise and quantify the joint diagnostic efficacy of this biomarker combination.

Our goal is to develop fully multivariate kernel estimators of ROC curves the construction of the ROC curve via a novel combination of kernel estimators of cumulative distribution and survival functions. In Section 2, we set up a framework for the squared error analysis of kernel estimators of multivariate cumulative distribution and survival functions, and most crucially the development of data-based optimal bandwidth selectors. This supporting section is the basis for the data-based implementation for the estimators of ROC curves in Section 3. In Section 4, we demonstrate the efficacy of these kernel estimators for simulated and experimental data. The last section is a discussion, and Appendix contains the mathematical proofs of the results stated in the main text.

2. Cumulative distribution and survival functions

In this section, we introduce a class of plug-in estimators of the cumulative distribution and survival functions. The usual kernel estimator of the cumulative distribution function F is

$$\hat{F}(\boldsymbol{x};\ \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) \tag{1}$$

where $\mathcal{K}(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} K(\boldsymbol{w})\, d\boldsymbol{w}$ for a multivariate kernel function K, the scaled integrated kernel is $\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}) = \mathcal{K}(\mathbf{H}^{-1/2}\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{w})\, d\boldsymbol{w}$, and $\mathbf{H}$ is the bandwidth matrix. This form as the integral of the classical kernel density estimator